Stakes in Testing: Not a Simple Dichotomy but a Profile of Consequences That Guides Needed Evidence of Measurement Quality

ETS RR–19-19

Richard J. Tannenbaum
Michael T. Kane

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Stakes in Testing: Not a Simple Dichotomy but a Profile of Consequences That Guides Needed Evidence of Measurement Quality

Richard J. Tannenbaum & Michael T. Kane

Educational Testing Service, Princeton, NJ

Testing programs are often classified as high or low stakes to indicate how stringently they need to be evaluated. However, in practice, this classification falls short. A high-stakes label is taken to imply that all indicators of measurement quality must meet high standards; whereas a low-stakes label is taken to imply the opposite. This approach can result in inappropriate allocation of resources and inadequate attention to needed evidence. We argue that "stakes" are better thought of as a profile of consequences. We suggest generalizable criteria for evaluating and responding to stakes in testing, with applications to licensure, employment, and K–12 accountability testing.

Keywords High stakes; low stakes; consequences; validity; licensure; accountability testing

doi:10.1002/ets2.12255

The distinction between high-stakes and low-stakes decisions plays an important role in designing and evaluating testing programs. It is generally accepted that testing programs that are used to make important decisions with serious consequences need to be evaluated against high psychometric standards (Geisinger, 2011; Mehrens & Popham, 1992; National Research Council, 1999; Plake, 2002, 2011). For example, in its chapter on the rights and responsibilities of test takers, the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) suggested the following:

> As with standards for the various phases of test development, when relevant standards are not met in test use, the reasons should be persuasive. The greater the potential impact on test takers, for good or ill, the greater the need to identify and satisfy the relevant standards. (p. 141)

The distinction between high and low stakes is intended to serve an important function by signaling when consequences of testing are more or less serious (for one or more groups of stakeholders) and, hence, when assurance of quality is more or less critical—for example, when an estimated reliability of .70 may be sufficient versus when an estimate of .90 is needed (e.g., Nunnally, 1975).

However, in practice, the simple dichotomous distinction has at least two limitations. First, in fact, there is no sharp distinction between high stakes and low stakes; the stakes or consequences vary along a continuum from very high stakes to very low stakes, and the perceived ordering of stakes tends to vary over stakeholders. The second limitation, which is the main focus of this paper, is that a simple dichotomous labeling may lead to an undifferentiated response to the concerns, or stakes, associated with the program. By this we mean that labeling a test as "high stakes" tends to be interpreted to mean that all basic indicators of measurement and testing quality (e.g., reliability, validity, scoring accuracy and consistency, accuracy of equating and scaling, fairness, test security) must meet high standards, whereas the label "low stakes" is taken to mean that not much, if any, evidence of measurement and testing quality is required.

This "all or nothing" dichotomy is not constructive. For high-stakes programs, it means that limited resources (staff, time, expertise, money) may not be properly allocated (i.e., spread too thin), and for low-stakes programs, it means that some important aspects of measurement and testing quality may inadvertently be placed in jeopardy; even low-stakes

Corresponding author: Richard J. Tannenbaum, E-mail: rtannenbaum@ets.org

programs can have some substantial consequences. We propose, instead, that stakes be thought of as a profile of consequences, rather than as a dichotomy, so that a differentiated response pattern occurs where attention can be focused on those aspects of measurement and testing quality that are most impactful (carry more consequence).

We begin by describing the basic differences between what are commonly referred to as high- and low-stakes testing programs (where the "program" includes tests, administration practices and policies, and procedures for generating and reporting scores that are to be used for certain purposes). We then introduce the notion of a profile of different kinds of stakes as a function of score use, testing conditions and context, and stakeholders. We conclude the paper with examples of how one might identify the stakes associated with a testing program and address measurement issues associated with these stakes.

Tests and test scores in and of themselves (thought of as *measures*) do not have stakes (Plake, 2011). It is only when the scores are used to make consequential decisions that the testing program, involving the testing procedures and the decisions based on the scores, can be said to have stakes.

It is clearly useful to think about the uses of test scores and the consequences associated with these uses in developing a testing program and in evaluating its effectiveness, but anyone who has tried to evaluate the probable consequences of a testing program with any precision is likely to recognize the difficulty of this task. However, it is not necessary to be able to unambiguously categorize a testing program as high or low stakes or to precisely rank programs in terms of their stakes for the evaluation of stakes to be useful. A relatively rough ordering of the overall stakes associated with the proposed uses of a testing program can be helpful in allocating resources, and a recognition of the different kinds of stakes inherent in a score use can draw attention to design issues that may need to get special attention.

Rather than simply classifying score-based decision programs as high or low stakes (or higher or lower stakes), we suggest that the consequences that are of most concern (and that give the score-based decisions their stakes) should be identified, and efforts should be made to address these consequences. In other words, we suggest that attention and resources be focused on those potential limitations of the testing program (e.g., modest reliability, sources of construct-irrelevant variance, gaps in test security) that could make negative consequences more likely or more severe or could make intended outcomes less likely. For example, licensure tests generally have high stakes for test takers, as failing to pass denies an opportunity to engage in professional practice and, thereby, can lead to economic and psychological hardship, but passing allows the candidate to practice his or her chosen profession (Downing & Haladyna, 1996). In this case, the stakes for the test taker and the public are concentrated almost exclusively on the accuracy of the pass/fail decision, and concerns about precision focus on decision consistency and on minimizing the standard error at the required passing score (AERA et al., 2014; Mehrens & Popham, 1992; Wyse & Babcock, 2016). In contrast, other high-stakes testing programs (e.g., college admissions tests) tend not to have fixed passing scores, and concerns about precision tend to emphasize the reliability of scores across a wide range of scores. Both licensure testing programs and admissions testing programs are considered high stakes, but the issues that merit extra attention because of the stakes are different in the two cases.

The General Distinctions

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), a *high-stakes test* is "a test used to provide results that have important, direct consequences for individuals, programs, or institutions involved in the testing" (p. 219), and a *low-stakes test* is "a test used to provide results that have only minor or indirect consequences for individuals, programs, or institutions involved in the testing" (p. 221). So the basic distinction involves a difference between "important, direct consequences," and "only minor or indirect consequences," for individuals (test takers), programs, or institutions. In the *Standards* (AERA et al., 2014), the role of stakes (or consequences) tends to be prominent in determining how specific standards are applied, for example, in the comments associated with particular standards (e.g., Standards 6.6, 6.13, 9.5, 12.1, 12.10, and 13.4) where stakes or consequences are invoked to indicate when a standard is especially important and needs to be fully implemented.

In describing a testing program as high or low stakes, the focus is on typical or expected outcomes. Consequences that affect particular test takers because of a unique personal situation or characteristic, but do not affect any substantial group of test takers, are not generally considered in categorizing testing programs as high or low stakes (but such unique situations are to be accommodated in particular cases, if appropriate and feasible). That is, the fact that we can imagine or know of a situation in which a score use could have a serious consequence in some unique, unusual circumstances does not make the testing program as a whole high stakes.

As noted earlier, consequences and therefore stakes are not associated with the test itself but, rather, are mainly attributable to the uses of (or decisions made from) test scores. The scores from a given testing program may be used for higher stakes decisions (e.g., admissions, promotions) or for lower stakes uses (e.g., feedback to students or teachers). Furthermore, a test score may have several uses, each with different potential consequences, and the different consequences may have different implications for the kinds of evidence needed to evaluate the program.

So, if a test initially designed to support low-stakes decisions is repurposed (Wendler & Powers, 2009) or retrofitted (Fulcher & Davidson, 2009) so that scores are used to make high-stakes decisions, the kinds of evidence needed to support that new, more consequential use, would need to be considered. And although consequences are associated with the use of test scores, the context and conditions of testing can moderate the impact of the consequences; therefore, they are also relevant to evaluating stakes. For example, a test that costs the test taker much more, in time or resources, to take than another test, may be perceived to carry higher stakes for that reason.

Consideration of stakes is further complicated by the fact that a testing program tends to impact multiple stakeholders, not just test takers (Geisinger, 2011; Lane, Parke, & Stone, 1998), and the stakes for a testing program are likely to be interpreted differently by different stakeholders (Lottridge, Winter, & Mugan, 2013; Plake, 2011). Depending on the testing purpose and intended score use, stakeholders may include students, job candidates, parents, schools, administrators, employers, and the larger community (Lane et al., 1998). An important example of stakeholder-moderated stakes occurs in the case of testing for K–12 accountability. The consequences for students (test takers) tend to be low in accountability programs, as the scores, for the most part, are used to make decisions about schools or teachers and do not have much, if any, direct impact on students (Geisinger, 2011; Koretz & Hamilton, 2006; Lane & Stone, 2002). However, the consequences for administrators and schools can be high, as failure of schools to meet defined standards of performance or improvement could result in their being restructured, converted into charter schools, or taken over by the state (Goertz & Duffy, 2003). As a result of the differences in perceived consequences (positive and negative) of state testing programs across various stakeholder groups, we have some groups of parents opting their children out of testing (Bennett, 2016).

Stakes: More Than a Simple Dichotomy

Just as validity is not a property of a test but of underlying assumptions, explicit claims, proposed interpretations, and intended uses of the test scores (Cronbach, 1988; Kane, 2013), stakes are not a fixed property of a test; rather, the stakes are inextricably tied to consequences of score use (e.g., Amrein & Berliner, 2002; Cizek, 2001; Cole & Osterlind, 2008; Geisinger, 2011). However, the use(s) of scores from testing programs tend to have different kinds of consequences, and consequences come in varying degrees and durations.

Furthermore, stakes associated with the same testing program tend to be evaluated differently by different stakeholders and can also depend on the conditions and context surrounding the testing (AERA et al., 2014; Bennett, 2016; Geisinger, 2011; Stone & Lane, 2003). Consider, for example, a relicensure program. Failing the test could result in a more extensive relicensure process, and it could interfere with the individual's current employment status. The impact of not passing the test is significant and consequential for test takers who fail; therefore, such programs are considered high stakes. However, these stakes would be even higher if the test is offered only once or twice a year and a nontrivial registration fee is required for each test occasion; if the test is offered more or less continuously and only one registration fee is required for a year's worth of testing opportunity, the stakes associated with the score use would generally be less serious. Although the consequences for the test taker of not getting relicensed have not changed, the first condition (infrequent testing and more cost) would make the consequences more serious than they would be in the second condition (continuous testing and less cost). The point is that classification of stakes is not a simple dichotomy of low or high, nor is it simply a function of the type of the test or even of the test-score use. There are many factors that determine the consequences, or stakes, associated with any testing program.

Rather than characterizing testing programs as high or low stakes, it would make sense to identify the specific kinds of stakes that are of most concern, to focus on actions that are likely to mitigate any negative consequences associated with the program, and if possible, to enhance the positive consequences. For example, consider Standard 6.6 in the *Standards*: "Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means" (AERA et al., 2014, p. 116). The comment associated with this standard emphasizes the importance of score integrity in cases "where the results may be viewed as having important consequences" (AERA et al., 2014, p. 116) but focuses on precautions against student cheating during testing (e.g., requiring adequate

space between students, monitoring the testing sessions, preventing communication with outside accomplices). It is certainly important to ensure the integrity and accuracy of scores in any high-stakes testing program, but the precautions that are most critical depend on the kinds of consequences that are of most concern. In contexts where the stakes are most salient for individual test takers (e.g., college admissions, licensure), focusing on test-taker cheating is appropriate; in contexts where the stakes are high for institutions (e.g., K – 12 accountability programs) but minimal to nonexistent for students (Stone & Lane, 2003), precautions against individual cheating should probably take a back seat to precautions against institutional malfeasance (e.g., inappropriate test preparation, tampering with scores after the fact). In thinking about how to design and conduct testing programs given the associated stakes, it is important first to identify the kinds of stakes at issue. In the next section, we identify four criteria that are relevant to evaluating the seriousness of consequences.

Criteria for Evaluating Consequences

Geisinger (2011) proposed several factors that determine the stakes associated with a testing program, including possible negative consequences for various stakeholders and various aspects of test use that influence the likelihood or severity of these consequences, such as (a) whether or not the test score is the sole source of information for making a decision, (b) the frequency with which a test is given and the opportunity for retaking the test, (c) how multiple scores earned by a test taker on the same test are considered (e.g., when the test taker repeats the test on different occasions is only the highest score counted, is an average of the scores counted or is the initial score weighted more heavily in the decision making?), (d) whether the scores earned are made public, and (e) the extent to which the test results are explained to test takers.

Geisinger (2011) associated higher stakes with programs for which a test score was the sole source of information but pointed out that this is, by and large, not considered a reasonable measurement practice (see, AERA et al., 2014). He associated higher stakes with programs (a) in which the tests are infrequently offered and for which retake opportunities are limited, (b) that average scores across multiple occasions or place disproportionate emphasis on the initial score, (c) for which scores are made public, and (d) that offered no or few explanations of test scores to the test takers.

Having explicit criteria against which to gauge the stakes of a testing program is important, as criteria tend to enhance the consistency with which programs are characterized in terms of stakes. Further, the identification and evaluation of the consequences of testing programs can help to identify specific negative consequences that need to be controlled through appropriate measurement practices and policy decisions. Geisinger (2011) offers a useful architecture for thinking about testing stakes, one that is oriented toward policy-based factors (e.g., whether test scores are made public). We offer a complementary set of criteria to facilitate the evaluation of stakes that may be somewhat more molar, yet are still actionable. We suggest four criteria: the direction of the consequences (negative vs. positive), the impact of the consequences, their likelihood, and their reversibility.

Direction of the Consequences

Testing programs resulting in only or mostly positive consequences for test takers (or other stakeholders) tend not to be subject to the same level of scrutiny as tests that can generate negative consequences. In evaluating the consequences of score use, doing no harm should be a primary concern (as suggested by the Hippocratic Oath). As Cronbach (1988) suggested: "Validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to argue against adverse consequences" (p. 6). In particular, concerns about stakes tend to occur mainly when the results are used to deprive some individuals of some opportunity or benefit, such as a license to practice a profession or occupation, admission to an institution, access to a high school diploma, or to impose some sanction (as in K – 12 accountability programs).

Two kinds of negative impact have been of particular concern in education: differential impact against particular groups (which may or may not be associated with identifiable sources of bias) and negative systemic effects (e.g., teaching to the test). These negative consequences are particularly salient because they have a clearly identifiable negative impact on clearly identifiable groups of students (i.e., the students who may have been treated unfairly and the students subjected to a limited curriculum). In addition, of course, these consequences can be serious, can have an impact on large numbers of students, and can be hard to correct after the fact. Negative consequences, especially specific, well-defined negative consequences, tend to be widely accepted as serious public concerns.

Although it is true that test takers are often the focus for discussions of negative consequences, testing programs tend to have stakes for multiple stakeholders, and so the extent to which a consequence is negative may depend on the particular stakeholder group of interest. For example, employment tests tend to have serious consequences for applicants who want to be offered a job and employers that want to hire the best applicants. K–12 accountability tests can carry high stakes for administrators and teachers, but they carry low stakes for students (Koretz & Hamilton, 2006; Stone & Lane, 2003).

Evaluating stakes across multiple stakeholder groups tends to be complicated, and the evaluation is not compensatory. Testing programs tend to be considered high stakes if they have serious consequences (especially serious negative consequences) for one or more stakeholder groups, even if the consequences for other stakeholder groups are neutral or somewhat positive.

Impact of the Consequences

A consequence is important to the extent that it tends to have significant impact on test takers, institutions, organizations, agencies, or the public directly affected by the test score use. Applying this criterion can be difficult because, as we have noted, testing programs can have a range of consequences for a range of stakeholders, and the evaluation of the importance of any consequence for any stakeholder is somewhat subjective. However, as we suggested earlier, in assigning stakes to testing programs, it is not necessary to get a precise ordering of the programs in terms of the relative importance of various consequences, and a rough ordering can be achieved.

For example, licensure testing outcomes have serious consequences for test takers because scores are used to decide eligibility for work in the test taker's chosen field. Test takers generally have dedicated several years of study and, in many cases, delayed income generation to prepare for a profession (e.g., physician, pharmacist, accountant) or skilled occupation (e.g., plumber, electrician, cosmetologist), and a failure to get a license would have a substantial negative impact on the test taker. A false-negative decision would be particularly problematic, because it would have no countervailing benefit in terms of protection of the public.

On the other hand, academic placement tests used to decide if incoming first-year university students may place out of beginning English composition would generally have less serious consequences. For placement, the downsides of a false-negative misclassification are taking one semester of beginning English and paying for the course; this is not a trivial consequence, but it is generally less serious than the consequences associated with a false-negative decision on a licensure examination.

Likelihood of Substantial Consequences

Consequences (especially negative consequences) that are more likely to occur are generally of greater concern than those that, although possible, are not probable. For example, if the same form of a multiple-choice test is administered repeatedly in a context with significant consequences, the items are likely to become known, and this type of security breach can undermine the meaning of the scores and the fairness of the decisions based on those scores; the high likelihood of this threat to fairness makes it a major concern in many cases. On the other hand, if students are encouraged to bring a basic calculator to a mathematics test, the possibility that a few students will forget to do so does not increase the stakes associated with the testing program. The suggestion that a negative consequence may occur under certain unusual circumstances does not necessarily raise the stakes associated with the score use.

Duration and Reversibility of Consequences

An outcome that is short-term or easily reversible tends to have lower stakes than an outcome of similar importance that is of extended duration or difficult (if not impossible) to reverse. In a K–6 context, a language proficiency test, for example, that incorrectly assigns a student to a lower language ability class may be corrected relatively quickly. A student-teacher whose first-week evaluation indicates a shortcoming with managing student behavior can readily take corrective action with the assistance of the supervising teacher. Educational assessments used formatively are intended to help students to acquire various knowledge, skills, and competencies, and their consequences are generally positive; any negative effects of the feedback are likely to be short-lived compared to the potential long-term benefits of the constructive feedback.

A false-positive licensure decision, however, tends to have long-lasting consequences; in many fields, once an individual is licensed to practice, it can be quite difficult to revoke the license. At the other extreme, tests that may be taken multiple times during a year, or, in cases of some Internet-based tests, continuously on demand, make it possible for test takers to achieve success after a false-negative decision much more readily than is the case for tests administered less frequently, thus minimizing the impact of the false-negative decision. Whether or not test takers should be permitted to retake a test as many times as they want until they pass is another matter, as the likelihood of a false-positive decision also increases with multiple opportunities to retake (Millman, 1989); such false positives can negatively impact the public if the test taker who eventually passes is poorly prepared for practice (Clauser & Nungester, 2001). A false-positive decision in an employment setting also may be long lasting and therefore costly to the employer.

Promoting Positive Consequences and Limiting Negative Consequences

As noted, identifying and evaluating the consequences associated with a score use tends to be a challenging and somewhat subjective endeavor; therefore, estimates of the stakes associated with a testing program are likely to be rough, rather than precise. The four criteria outlined previously (see also Geisinger, 2011) can be useful in evaluating the overall level of stakes for a testing program, but probably more importantly, they can help to identify specific consequences that contribute to the stakes associated with the program. To the extent that some aspect of the program is likely to generate serious negative consequences or to seriously interfere with the obtainment of intended benefits, these aspects need to be addressed.

The distinction between higher stakes and lower stakes programs is useful, but labeling a program as high or low stakes does not, in itself, provide much guidance on how to improve the program. Given that high-stakes testing programs can have different profiles of consequence, it would probably be more useful to identify the specific consequences that are of concern and to address these consequences directly rather than simply to categorize some programs as high stakes and thereby require a high level of attention and scrutiny on all aspects of measurement-related quality while allowing all other programs to escape such scrutiny.

An effective way to deal with questions about stakes, consistent with the notion that each test has a definable profile of consequences, is to spell out the proposed interpretation and intended uses of the test scores (Kane, 2013), including the likely consequences of the proposed uses and of the testing conditions and context (e.g., retesting policies). Based on these analyses, the most important consequences, positive and negative, could be identified, and efforts to improve the quality of the assessments could be focused on enhancing the positive consequences and controlling the negative consequences.

For example, in a licensure testing program, false-positive decisions (granting a licensure to an individual not ready to engage in practice) and false-negative decisions (not granting a licensure when it is merited) could have serious negative consequences for the public and for the candidate; therefore, the consistency and accuracy of the decisions need to be high. This need implies that the standard-setting procedures followed to identify the passing score must be appropriate and rigorous, and score precision should be maximized around the passing score. The precision of scores around the passing score can be enhanced during the test-development process by including a larger number of items with difficulty levels close to the likely passing score rather than at higher and lower levels. Although as Wyse and Babcock (2016) recently noted, depending on the length of the test and the distribution of test-taker ability relative to the passing score, decision accuracy and consistency may, in fact, be maximized by including items above or below the passing score. In the case of constructed-response tasks, the need to maximize decision consistency could suggest, for example, the rescoring of all responses within a defined band around the passing score.

In K–12 accountability programs, negative consequences that can be anticipated are a narrowing of the curriculum and instruction to emphasize the content covered by the test (Au, 2007; Koretz & Hamilton, 2006; Lane et al., 1998; Lane & Stone, 2002). Under these circumstances, it is particularly important that the test cover the entire curriculum, or at least the most important parts of the curriculum (AERA et al., 2014; Bandalos, Ferster, Davis, & Samuelson, 2011; Elliott, 2015). By way of contrast, there is less evidence indicating that licensure tests have much impact on the curricula and instructional practices of professional schools (Faxon-Mills, Hamilton, Rudnick, & Stecher, 2013), and licensure tests do not need to cover all occupationally relevant knowledge, skills, and judgment (KSJ; AERA et al., 2014).

Thinking Through Consequences: Examples of Testing Applications

In this section, we offer some examples of how one might consider consequences in the context of three testing applications. First, we will examine some of the consequences (or stakes) associated with the three uses of test scores, licensure decisions, employment decisions, and K–12 accountability. Second, we will consider the implication of these consequences for the level of rigor needed for selected aspects of measurement quality for these applications.

Evaluating Stakes for Licensure Testing

Licensure is a strict form of regulation, as individuals cannot legally perform the primary responsibilities of the profession or occupation without first obtaining a license (AERA et al., 2014, ch. 11; Shimberg, 1981). The intent of licensure is to minimize the likelihood that unqualified individuals enter a profession or occupation. Schmitt (1995) noted, "Licensure is designed to protect citizens from mental, physical, or economic harm that could be caused by practitioners who may not be sufficiently competent to enter the profession" (p. 4). A license tends to apply to a broad field of practice (e.g., architecture, accounting, teaching) in a range of contexts rather than to a specific job within that field. Licensure tests evaluate current mastery of KSJs needed for the practice of a profession or skilled occupation; they are not designed to predict future performance (Kane, 2004; Schmitt, 1995).

> If a [licensure] test is developed to assess the extent to which professionals possess the skills required for practice, it is sensible that examinees with lower levels of those skills will be less-fit practitioners. But examinees who possess these skills may fail to use them; knowing how and doing what is required are two different things. (Clauser, Margolis, & Case, 2006, p. 716)

In discussing licensure and other credentialing examinations, the *Standards* (AERA et al., 2014) stated:

> Criterion-related evidence is of limited applicability because credentialing examinations are not intended to predict individual performance in a specific job but rather to provide evidence that candidates have acquired the knowledge, skills, and judgment required for effective performance, often in a wide variety of jobs or settings. (pp. 175–176)

In licensure testing, the passing score provides the operational basis for identifying candidates who have the KSJs required for entry-level practice. Candidates with scores meeting or exceeding the passing score are allowed to practice; those below the passing score are considered not yet ready to enter the profession. The percentage of candidates passing a licensure test will vary from administration to administration, and in principle, the passing percentage could be as low as 0% or as high as 100%. For licensure programs, serious negative consequences may occur, as noted previously, both for the public and test takers, and so from either of these stakeholder perspectives, licensure tests carry high stakes.

Evaluating Stakes for Employment Testing

Employment (selection) testing seeks to identify applicants who will be most successful on the job for which they are being hired. Unlike licensure testing, which is intended to weed out inadequately prepared candidates, "the fundamental inference to be drawn from test scores in most applications of testing in employment settings is one of prediction: The test user wishes to make an inference from test results to some future job behavior or job outcome" (AERA et al., 2014, p. 171).

In employment testing, applicants are typically rank-ordered from highest scoring to lowest scoring on the selection test (or on multiple measures) with the highest scoring applicants being offered jobs or, at least, included in the next stage of the selection process. The number of actual hires is determined, in large part, by the number of open positions, and there is no fixed passing score. We recognize that employment testing covers more than selection (hiring) decisions and includes, for example, decisions related to placement, training, and promotion. However, for present purposes, we focused on selection.

Employment testing decisions also have high stakes, but generally, the stakes are not as high as those associated with licensure tests. The institutional impact of any employment testing program may be significant, but it tends to be limited to the organization doing the hiring. For the applicants who take the test, not getting a job they want can have serious

consequences, but the impact of negative decisions is mitigated in most cases by opportunities to immediately apply for other jobs.

Evaluating Stakes for K–12 Accountability Testing

The third example is K–12 accountability testing, which has become widespread in U.S. public schools (Coburn, Hill, & Spillane, 2016; Koretz & Hamilton, 2006). Such testing is often focused on gauging students' mastery of the state-mandated content curriculum with the goals of identifying where instructional and school-level changes may be most needed to support student learning (Bandalos et al., 2011; Lane, 2014). Although students, in the aggregate, are the focus of the assessment process, the stakes associated with their test scores tend to be higher for administrators and teachers, especially when rewards and sanctions are attached to school-level performance (Lane & Stone, 2002). However, students may be affected by unintended consequences, including a narrowing of the curriculum and instruction to the topics covered by the state-required test (Koretz & Hamilton, 2006; Lane et al., 1998; Lane & Stone, 2002). Although better alignment between the curriculum and the test could improve instructional relevance (Woolley, Rose, Orthner, Akos, & Jones-Sanpei, 2013), state testing programs do not generally cover the full range of competencies included in the state content standards.

All three of these kinds of testing programs can be considered high stakes in that they each tend to have serious negative consequences for at least some stakeholders (see Table 1). Licensure and employment tests have particularly direct, serious consequences for test takers, while K–12 accountability programs may have direct and serious consequences for administrators and teachers and indirect consequences for students. Employment tests can also have serious consequences for employers, and licensure tests can have negative consequences for the public and institutions that rely on the competence of professional practitioners. Note, however, that although all three kinds of testing programs tend to have serious stakes for various groups of stakeholders, the stakes are different in the three cases and the implications for the quality and rigor of the testing program are different, and so the simple label of "high stakes" does not provide adequate guidance for evaluating and improving the programs. Working through responses to the criteria (and the associated assumptions) can help to inform subsequent decisions about the evidence needed to evaluate the program.

The three kinds of testing programs discussed here also differ on the duration and reversibility of the consequences of the decisions being made. A false-negative outcome may be addressed by retaking the licensure test (most likely an alternate form of the test), but how quickly that may occur depends on the retake policy. An applicant who is not hired for a particular job can immediately apply for different jobs, assuming that comparable positions are available. The consequences of not getting a job increases if comparable opportunities are limited. An individual student may not be impacted by a school not meeting targeted expectations in an accountability program, but students in the aggregate, over time, may be subjected to a more restricted curriculum, and some students may be less motivated to learn (Lane & Stone, 2002).

Although the discussion of these three kinds of programs, summarized in Table 1, may suggest a possible rank ordering of the stakes associated with these three testing applications, such an ordering is not too critical; all three kinds of programs have serious enough potential consequences for various stakeholders to be considered high stakes. More important, in each case, are the steps that need to be taken to promote the positive consequences associate with each program and to limit any negative consequences associated with each program.

Addressing the Consequences That Give Testing Programs Their Stakes

As noted above, categorizing testing programs as high or low stakes or rank-ordering the programs in terms of their stakes is useful in thinking about how much attention to give to the technical quality of each program, but it does not provide much guidance on how much emphasis to give to specific technical criteria. To identify the technical issues that require the most serious attention, it is necessary to analyze the profile of consequences associated with the program and the technical issues that have an impact on these consequences.

Table 2 emphasizes the test-design and implementation issues that are highlighted by the stakes associated with licensure, employment, and K–12 accountability decisions and the kinds of evidence needed to address how well these issues are being addressed (but it does not reflect all of the criteria that may be of interest).

Note that the relevance of any kind of consequences for any particular testing program will depend on the intended uses of the scores and the context in which the program operates; our discussion of how consequences play out in each

Table 1 Evaluation Criteria Applied to Stakes (Consequences)

Evaluation criteria	Licensure testing	Employment testing	Accountability testing
Direction of the consequences *The use of the test may lead to negative consequences for one or more stakeholders.*	Candidates (test takers) and the public may experience direct negative consequences.	Applicants (test takers) and hiring organizations may experience direct negative consequences.	Schools (as a collective of administrators and teachers) may experience direct negative consequences.
Importance of the consequences *The impact of the consequences on stakeholders.*	Test takers who fail to pass are denied eligibility to practice their desired profession. A false-positive decision, may place the public in "harm's way."	Test scores, generally are only one criterion in the hiring decision, but test takers who fail to pass may be denied access to a job. A false-positive decision, may result in productivity loss and reduced morale for the organization.	Schools may face sanctions, if accountability metrics are not met. Students may face a narrowing of the curriculum and related instruction.
Likelihood of important consequences *The probability that negative consequences will occur.*	Inaccurate decisions are most likely to occur for test takers and the public with true scores near the passing score.	Inaccurate decisions tend to occur if the test scores are not strongly related to job performance.	Schools already facing resource and funding issues may be more prone to negative consequences than other schools.
Duration and reversibility of negative consequences	A test taker who fails to pass may retake the licensure test, but this may take months. It is difficult to revoke a license once granted.	A test taker who is not hired for a job can immediately apply for other jobs, if available.	Meeting accountability metrics may be difficult, given available resources; and positive change may take a long time to develop.

kind of testing program is general and tentative. The basic point is that it is possible to identify the issues that need to be addressed in order to enhance intended, positive consequences and to minimize the impact of negative consequences.

The definition of the content domain to be covered by the test and the extent to which the test content matches this target content domain is of importance in any testing program—hence, in Table 2, we have coded "domain representativeness" as being of high importance across the three score-use applications—but the specificity of the domain definition, and the methods used to articulate its structure and boundaries, depends on the intended interpretation and use of the scores. Licensure programs are intended to protect the public by weeding out candidates who have not mastered the KSJs that entering practitioners need for safe and effective practice. Domain representativeness tends to be an especially important consideration in the development and evaluation of licensure tests (AERA et al., 2014) because the test scores are interpreted primarily in terms of the level of mastery of these job-relevant KSJs (as determined by a job or practice analysis) and covered by the test. For a licensure examination, it is not necessary to include all of the KSJs needed for practice, but it is important that the KSJs assessed by the examination be relevant to safe and effective practice (AERA et al., 2014).

For employment tests and K–12 accountability tests, domain representativeness is also of importance, especially in defining the constructs to be measured by the test. In the former case, the content is intended to reflect general characteristics needed for success in a particular job or in training programs, but specific KSJ domains are not necessarily the focus, although there needs to be an evidence-based connection (e.g., through a job analysis) between the tested content and job requirements.

In K–12 accountability testing, alignment between the tested content and the state-approved curriculum or content standards is important (Bandalos et al., 2011; Davis-Becker & Buckendahl, 2013; Koretz & Hamilton, 2006; Webb, 1999). Evidence of such alignment, at a minimum, focuses on the overlap between the knowledge and skills represented on the test and the state curriculum or standards and on the extent to which the tested content reasonably reflects the level of cognitive complexity depicted by the curriculum or standards (Bhola, Impara, & Buckendahl, 2003; Webb, 2007). For K–12 accountability programs, it is necessary to cover the state approved standards, because the kinds of performance

Table 2 How the Level or Kind of Evidence Differs Across the Three Score-Use Applications

Measurement criteria	Licensure testing	Employment testing	Accountability testing
Domain representativeness	(H) Test content should reflect the specific knowledge, skills, and judgments (KSJs) needed for safe and effective practice.	(H) Test content should reflect characteristics and KSJs needed for success on the job or in training for the job.	(H) Test content should reflect the knowledge and skills defining the state-mandated curriculum or standards.
Criterion-related validity	(L) Focus is on the extent to which test scores reflect mastery of the KSJs needed in practice rather than on criterion-related validity evidence.	(H) The main validity concern is the relationship between test scores and performance on the job.	(L)The main concern is evidence of the alignment between the test content and the state-mandated curriculum or standards.
Precision/Reliability of individual scores	(M) Precision should be high around the passing score; less precision is acceptable at other score scale locations.	(H) Scores should be reliable over the range of scores in which decisions might be made.	(L) The focus is on aggregate results for the unit being evaluated, rather than on individual reliability.
Fairness/Lack of bias	(H) Tested content should not advantage (disadvantage) one subgroup over another.	(H) Testing practices should not discriminate against applicants based on job-irrelevant factors such as race, sex, national origin.	(H) Administrators and teachers should not be accountable for students' test scores without clear guidelines on what is to be taught and without the resources needed to provide effective instruction.

Note. We classified the evidence in the table as either high (H), moderate (M), or low (L) importance, mostly as a way to illustrate how we believe the same source of evidence may or may not vary across score-use applications.

that are not assessed by the testing program are likely to get less instructional attention than performances that are included.

All three kinds of testing programs need to be supported by validity arguments, but the mix of evidence needed to support the validity of the proposed interpretation and use is not the same in the three cases. As noted above, content-related evidence is important for all three kinds of interpretations and uses. A second kind of validity evidence that plays a major role in some programs is criterion-related evidence (Table 2, second row). Validity arguments for licensure testing tend to focus primarily on domain representativeness (job-analytic evidence) and on the reasonableness of the passing score (standard-setting evidence), and as indicated earlier, there is no expectation that the test scores will provide predictions of the future performance of individual licensees; hence we coded "criterion-relation validity" as being of low importance for licensure testing in Table 2. For employment testing, validity arguments rely heavily on criterion-related, predictive evidence (particularly, the empirical relationship between test scores and criteria of job performance such as supervisor ratings), and so we coded criterion-related validity as being of high importance. Validity arguments for K–12 accountability testing focus on the alignment of test content to the state curriculum or standards (U.S. Department of Education, 2004); so as in the case of licensure testing, "criterion-related validity" is of low importance for accountability testing.

The third row in Table 2 compares the different needs for precision or reliability of individual test scores in the three kinds of testing programs. Employment tests seek to rank-order applicants in terms of their predicted performance on the job and need scores that are reliable over much of their score range; hence, evidence of precision or reliability is of high importance. Licensure tests need to generate consistent pass/fail decisions, and in order to minimize the likelihood of false positives and false negatives, they need to have relatively small standard errors around the passing score; reliability across the range of the score scale is less of a concern. We, therefore, have classified the need for precision or reliability as of moderate importance. For K–12 accountability, where scores are interpreted at a class or school level (cohorts of students) rather than at an individual student level, it is the reliability of class or school means that is the main concern in

evaluating precision, and the reliability of individual student scores is not a major concern; hence, we coded such evidence as being of low importance. If the accountability decisions are based on the numbers of students in different performance levels (e.g., basic, proficient, advanced), it is desirable that precision be high around the cut scores to enhance classification accuracy.

The last row in Table 2 addresses the issue of fairness. A test is said to be fair if it "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population" (AERA, APA, NCME, 2014, p. 50). A fair test does not mean that different groups of test takers will perform comparably on that test, but it does mean that observed differences in performance, on average, are not due to factors unrelated to what the test was intended to measure. We see fairness in testing as nonnegotiable, and so we coded it as being of high importance across the three score-use applications.

Fairness in employment testing (for hiring decisions) is, in large part, governed by the Civil Rights Act of 1964, which makes it clear that employers may not refuse to hire (or otherwise discriminate against) applicants due to their race, color, religion, sex, or national origin. Further, disparate passing rates (adverse impact) between test-taker subgroups is taken as an indicator of potential bias, where adverse impact is said to occur if the passing rate (selection rate) for some subgroup of minority test takers is less than 80% of the rate for majority test takers (the so-called 4/5ths rule, e.g., Roth, Bobko, & Switzer, 2006). In such instances, the test user is required to provide evidence that the test scores are predictive of job performance.

Fairness in the context of licensure tends to focus on assuring that neither the test content nor its format and administration systematically advantages (disadvantages) one subgroup of test takers over others. If the program is found to have adverse impact, additional scrutiny will be given to the extent to which the test tasks reflect KSJs that are necessary for safe and effective practice and the possibility that sources of irrelevant score variance are introducing bias.

Lastly, in K – 12 accountability testing as such, where there are no direct consequences for individual students, concerns about fairness at the student level tend to be low, but concerns about fairness for the units (e.g., classes, schools) being evaluated for accountability are major concerns.

If the scores are also used to make decisions about individual students, we have an aspect of fairness not encountered in licensure and employment settings, which is the issue of opportunity to learn (Kurz, Elliott, Kettler, & Yel, 2014). This issue relates to the students having been provided adequate access to materials and instruction on the tested content and instructors having information about the content to be covered well in advance of the assessment administration. Elliott (2015) reflected that opportunity to learn includes the time the teacher devotes to instruction that covers different parts of the content in the state-mandated curriculum. The basic concern is that schools, administrators, and teachers should not be accountable for student performance without clear guidelines on what is to be taught and without the resources needed to achieve these educational goals.

As noted earlier, it is not particularly important to rank-order testing programs in terms of stakes. Whether a score use carries higher stakes or lower stakes, we want the testing program to work well—to achieve its intended objectives. So it is probably less important to determine how high the stakes are and more important to determine what the stakes are and, in particular, to determine how to promote intended positive consequences and to limit or mitigate any unintended negative consequences.

Conclusion

The use of the terms, high stakes and low stakes, when referring to a test score use is commonplace. Such labeling is a convenient way of thinking about and discussing tests and testing programs. However, this distinction is, at best, rough and can be misleading. In particular, this simple dichotomy may communicate a set of assumptions about the level of measurement rigor needed that is not adequately differentiated. To the extent that the label "high stakes" invokes an almost automatic, generalized response where all aspects of measurement-related quality (e.g., security, validity, reliability) must be extensive and premium, it can lead to over-emphasizing certain indices and evidence and ineffective use of limited resources (e.g., time, expertise, finances). A more informative approach would be to think about a test score use as having a profile of stakes (or consequences) and identifying positive consequences to be promoted and significant (impactful) negative consequences that need to be controlled or mitigated.

In this paper, we have offered some criteria for evaluating stakes and suggested how careful attention to the delineation of explicit claims, uses, and especially consequences may be helpful in promoting desirable outcomes and in controlling

unintended, negative consequences. We further highlighted how testing context and conditions moderate stakes and so must be taken into account, and we similarly reinforced the need to take into account all stakeholders when thinking about stakes.

We suggest that the particular consequences that give a high-stakes program its stakes should be identified, and particularly rigorous standards should be adopted for those aspects of the testing program that are most critical for promoting positive consequences or for mitigating probable negative consequences. Our point is that the level and kinds of evidence needed to support a testing program, or more specifically, the use of scores from that program, should be commensurate with and reflect the profile of consequences associated with that use. A lower stakes test may not need the same level of rigor as a higher stakes test in most areas, but it may have a need for more evidence in some areas, and a higher stakes program will generally be held to a higher standard than lower stakes programs, but it may not need as much attention in some particular areas as a lower stakes program. In both cases, the issues that determine a program's stakes need to be addressed by focusing on those aspects of the testing program that are most directly related to its stakes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*, 1–74. https://doi.org/10.14507/epaa.v10n18.2002

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*, 258–267. https://doi.org/10.3102/0013189X07306523

Bandalos, D. L., Ferster, A. E., Davis, S. L., & Samuelson, K. M. (2011). Validity arguments for high-stakes testing and accountability systems. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 155–175). Washington, DC: American Psychological Association. https://doi.org/10.1037/12330-010

Bennett, R. E. (2016). *Opt out: An examination of issues* (Research Report No. RR-16-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12101

Bhola, D., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, *22*(3), 21–29. https://doi.org/10.1111/j.1745-3992.2003.tb00134.x

Civil Rights Act of 1964 § 7, 42 U.S.C. §2000e *et seq*. (1964). Retrieved from http://www.eeoc.gov/laws/statutes/titlevii.cfm

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, *20*(4), 19–27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–731). Westport, CT: Praeger.

Clauser, B. E., & Nungester, R. J. (2001). Classification accuracy for tests that allow retakes. *Academic Medicine*, *76*, S108–S110. https://doi.org/10.1097/00001888-200110001-00036

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*, 243–251. https://doi.org/10.3102/0013189X16651080

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*, 119–130. https://doi.org/10.1353/jge.0.0018

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement: Issues and Practice*, *32*(1), 23–33. https://doi.org/10.1111/emip.12002

Downing, S. M., & Haladyna, T. M. (1996). A model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practice*, *15*(1), 5–12. https://doi.org/10.1111/j.1745-3992.1996.tb00801.x

Elliott, S. N. (2015). Measuring opportunity to learn and achievement growth key research issues with implications for the effective education of all students. *Remedial and Special Education*, *36*, 58–64. https://doi.org/10.1177/0741932514551282

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement*. Santa Monica, CA: Rand.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, *26*, 123–144. https://doi.org/10.1177/0265532208097339

Geisinger, K. F. (2011). The future of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 231–248), Washington, DC: American Psychological Association. https://doi.org/10.1037/12330-014

Goertz, M. E., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory Into Practice*, *42*, 4–11. https://doi.org/10.1207/s15430421tip4201_2

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 135–170. https://doi.org/10.1111/jedm.12000

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Kurz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment*, *19*, 159–184. https://doi.org/10.1080/10627197.2014.934606

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 24–27. https://doi.org/10.1111/j.1745-3992.1998.tb00830.x

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment for accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23–30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Lottridge, S., Winter, P., & Mugan, L. (2013). *The AS decision matrix: Using program stakes and item type to make informed decisions about automated scoring implementations* (Research Publications). Monterey, CA: Pacific Metrics Corporation.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, *5*, 265–283. https://doi.org/10.1207/s15324818ame0503_5

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Research*, *18*(6), 5–9.

National Research Council. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Nunnally, J. (1975). Psychometric theory—25 years ago and now. *Educational Researcher*, *4*(10), 7–14, 19–20.

Plake, B. S. (2002). Evaluating the technical quality of educational tests used for high-stakes decisions. *Measurement and Evaluation in Counseling and Development*, *35*, 144–152.

Plake, B. S. (2011). Current state of high-stakes testing in education. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K–12 settings* (pp. 11–26), Washington, DC: American Psychological Association. https://doi.org/10.1037/12330-001

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology*, *91*, 507–522. https://doi.org/10.1037/0021-9010.91.3.507

Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Lincoln, NE: Buros Institute of Mental Measurements.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36,* 1138–1146. https://doi.org/10.1037/0003-066X.36.10.1138

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education*, *16*, 1–26. https://doi.org/10.1207/S15324818AME1601_1

U.S. Department of Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Office of Elementary and Secondary Education.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Madison: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*, 7–25. https://doi.org/10.1080/08957340709336728

Wendler, C., & Powers, D. (2009). What does it mean to repurpose a test? *R & D Connections*, *9*, 1–8.

Woolley, M. E., Rose, R. A., Orthner, D. K., Akos, P. T., & Jones-Sanpei, H. (2013). Advancing academic achievement through career relevance in the middle grades: A longitudinal evaluation of CareerStart. *American Educational Research Journal, 50*, 1309–1335. https://doi.org/10.3102/0002831213488818

Wyse, A. E., & Babcock, B. (2016). Does maximizing information at the cut score always maximize classification accuracy and consistency? *Journal of Educational Measurement*, *53*, 23–44. https://doi.org/10.1111/jedm.12099

Suggested citation:

Tannenbaum, R. J., & Kane, M. T. (2019). *Stakes in testing: Not a simple dichotomy but a profile of consequences that guides needed evidence of measurement quality* (Research Report No. RR-19-19). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12255

Action Editor: Rebecca Zwick

Reviewers: Larry Stricker and Michael Zieky